Rejoinder

Bart De Ketelaere[*], Eric Schmitt[**], Tiago Rato[*], Mia Hubert[**]

Leuven Statistics Research Centre, KU Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium

[*]Department of Biosystems, Division MeBioS, KU Leuven, Kasteelpark Arenberg 30, B-3001 Heverlee, Belgium

[**]Department of Mathematics, KU Leuven, Celestijnenlaan 200B, B-3001 Heverlee, Belgium

We would like to thank Giovanna Capizzi, Marit Schoonhoven and Rob Goedhart for their discussion papers, and we are thankful for having the opportunity for replying on these. Both have replied with interesting and constructive insights on the challenges of complex process monitoring. In our closing comments, we will address their main points, which we split up into 5 categories. The first one relates to a common thread throughout both papers, being the choice between data driven models, which we advocated, and model driven approaches as alternatives. The second discussion point is the use of latent variables techniques such as PCA for sparse data, and the issue of how well those techniques perform for detecting various types of out of control situations, which was a topic which we purposely excluded from our paper. Next, we will expand on the specific simulation settings we have chosen. The fourth section will handle the fair point that difference must be made between time-dependent measurements and time-dependent processes. Lastly, we address the issues raised when discussing the cointegration approach.

Data-driven models versus model-based approaches

The latent variable methods we described are not always easily interpretable in terms of their physical meaning, which can be a drawback. However, they can provide useful insights towards developing an engineering model, or at the very least a transparent expression of important relationships in the process. Both discussants raise the point that purely statistical models can be suboptimal because they ignore knowledge about the structure of the process and alternatives that do take the advantage of such process knowledge are to be preferred.

In engineering, first-principle models are derived for processes when possible. When such models are available and accurately describe the underlying process, they must indeed not be overlooked. Their use will typically lead to more powerful and interpretable monitoring schemes compared to purely data-driven methods because they are incorporating structural knowledge into the statistical model. We find that especially the interpretability of the models will greatly benefit from including process knowledge. As pointed out by Capizzi, this is especially the case when analyzing more complex data, such as multiway tensor data. Besides mixed and multilevel models, also multiway models (PARAFAC, Tucker3,…) can improve the online monitoring of those type of data (Meng et al., 2003).

One of the directions of research is blending model and data based approaches, which is a challenging topic. A particular challenge to developing blended models is that structural aspects of processes are often incompletely understood. Our experience is that practitioners have some insights into the causal mechanisms driving their processes, but subtle characteristics are often only

understood intuitively, or are not identified at all. Structural knowledge and latent variable methods can be integrated by making specific adaptations to the latent variable model. So-called “grey box” models that blend data-driven and model-based methods were for instance proposed by Gurden et al. (2001), where they describe the modelling of spectroscopic batch process data using grey models to incorporate external information. This work was further elaborated by Westerhuis et al. (2007) where Grey Component Analysis (GCA) is introduced as a term to denote a latent variable approach that is blended with physical models. Related, the blending of latent variables models with functional data is promising as well, as profile monitoring (and, by extension image monitoring) is a topic that gains attention (Noorossana et al., 2011).

Even with the advent of methods using engineering and statistical knowledge, such as those mentioned above and others that are mentioned in the discussion papers, the models produced will still fall short of perfectly explaining the processes under study (“*all models are wrong but some are useful*”). Our paper explores the consequences of imperfect modeling in the PCA context and we believe that these insights and some of the corrections we discuss will be relevant whenever the modelling exercise does not produce perfect results, which we suspect will be rather often.

Latent variable approach

Next, we turn to comments on the applicability of latent variable methods. The main goal of our discussion on PCA-based process monitoring of time-dependent data focusses on its capacity for describing the underlying data well, and not on fault detection capability. We fully agree, and also mentioned in the paper, that this is only a part of the story and a concise analysis of fault detection is essential. This last issue forms the topic of a recent paper which provides insight in how those methods do behave under AR and ARI processes, as well as their moving average counterparts (Rato et al., 2015). Based on an extensive simulation study we deem it essential that a monitoring scheme should describe the underlying process adequately – if not, fault detection will be impeded and false detection rates (FDRs) will be off-target.

The chosen simulation settings cover the high dimensional case (‘multivariate’), but not the case where dimensionality is very high such as in spectral applications (‘megavariate’). This need does not pose a problem for PCA methods, which are capable of handling cases where p >> n and are widely used in chemometrics. These methods are actually more sensitive to the number of latent variables than the number of variables in the data space. However, if there are many unimportant variables, or variables that are not important in some of the loadings, then sparse methods may be useful. By accelerating the “zeroing” of the influence of parameters on loadings where they are not relevant, model accuracy and interpretability can be increased. The suggestion that sparse PCA may be a useful process monitoring tool in this context is plausible. Sparse PCA for monitoring purposes is still not very much explored, but we are aware of sparse, robust PCA methods that find fits for the in-control data that are capable of revealing faults (see for example Hubert et al., 2015). Such extensions would improve interpretability, making these methods more competitive with solutions informed by engineering. The addition of the notion of robustness makes this method an appealing one for process monitoring, especially in Phase I for non-adaptive models. Eventually it might also be useful for the adaptive methods where updating in a robust scheme could be an alternative for the current practice, which is ignoring the out of control points and stacking the first in-control point after an out of control situation to the latest in-control point before this, inherently destroying the underlying dynamics. A different point of view is considered in Xie et al. (2013). Here, sparsity is introduced in the residual space to improve the interpretability of the faults.

Simulation settings

A number of points were mentioned in the Discussion papers regarding the dimensionality of the simulations we performed and the criterion we used for specifying the models we fit. We address these points in turn.

In our simulation settings we started from a 5 dimensional subspace and transformed them into a dataset of 50 variables. At 50 dimensions, many classical time-series methods cannot be applied. Therefore, an approach based on latent variables, or sparsity or ridging approaches is useful. We also feel this is an acceptable approximation of a wide range of contemporary processes, although we realize that the typical dimensionality increases with time posing additional issues.

Whatever the dimensionality, a criterion must be chosen to determine the number of underlying components. This was done in our paper using the Cumulative Percentage of Variation (CPV) which is explained by the model. The exact value of the CPV does certainly influence the monitoring results – when a smaller CPV is used, the model explains less variation in the data. This unexplained variation is transferred from the T^2–statistic to the Q-statistic, and the smaller the CPV value is set, the more the Q-statistic will behave like the T^2–statistic. The CPV was taken as constant throughout our paper to eliminate yet another potential influencing factor.

Given the dimensions of the original variables and the subspace, we have chosen to also fix the autocorrelation structure at the subspace level to a fixed value for all five dimension. As pointed out by Capizzi this is a restricted scope which will probably not be the case in practice, but narrowing the scope is necessary because interpretation is difficult even in this narrow case. The fact that we only used first order models (AR(1) and ARI(1,1)) is partly motivated by the same reasoning, but also by the fact that such models approximate a wide range of processes well.

Time-dependent measurements vs processes

In their paper, Schoonhoven and Goedhart point to the distinction between time-dependent measurements and time dependent data. This is a fair point, and both need to be tackled with different approaches, with time-dependent processes being the most challenging case.

Schoonhoven and Goedhart point out that in cases where the autocorrelation on the measurements is due to sampling speed, or is not severe, the monitoring statistics may be autocorrelated, but this should not have a major impact on fault detection. For time-dependent measurements, which we label as stationary but autocorrelated, we concluded that there are indeed no big issues as long as the dynamics are not too strong. This agrees with previous researchers and several approaches are advocated, ranging from ‘ignoring simple dynamics’ (Wheeler, 1991) to ‘applying control charts to residuals’ (Alwan and Roberts, 1988). For higher autocorrelations we feel the time-dependency should not be overlooked, confirming recent results of Vanhatalo and Kulahci (2015) who state that ignoring autocorrelation and using theoretical UCLs can lead to wrong conclusions, with in-control ARLs different from their nominal value. Such strong autocorrelations do not only appear when sampling very rapidly, but are also often visible in processes under closed loop control. In such cases, variability in FDR is expected across time which is undesirable.

The time-dependency of processes, which we label as nonstationary and autocorrelated, require a procedure that can actually model this process across time. We compared several such models in our paper but conclude that their implementation is not straightforward because of the issues around choosing their appropriate tuning (forgetting) parameters. Furthermore, although an ARI(1,1) model might seem a relatively simple case, its dynamics are complex and play at different frequencies – low

frequencies related to the overall dynamics of the process, and higher frequencies that relate to the autocorrelation structure. It should be further investigated how those dynamics can best be incorporated into the monitoring schemes. In light of the limitations on perfectly modelling a process such as the ARI(1,1), we propose to adjust the control limits to account for violations of the distributional assumptions made on the monitoring statistics to achieve FDR values that are in line with expectations while being able to carry out fault detection.

To conclude this section, we feel that despite the weaknesses we see, the general applicability and decency of the adaptive methods still makes them relevant tools for practitioners.

Cointegration

In our paper we introduced cointegration and opened the discussion whether it is a valuable candidate for monitoring nonstationary processes, because it naturally fits that kind of processes.

The answer to the question whether "*we are sure that cointegration is the most appropriate and feasible tool for monitoring nonstationary processes in the high dimensional framework*" simply is no. It has, to our knowledge, never been tested in detail for such situations. As mentioned in our paper and by Capizzi's response it is clear that the method, when applied in its basic form, does face challenges that need to be resolved when applying it to the high dimensional case, especially because the number of parameters to be estimated grows rapidly with increasing dimensionality, and the fact that tests for determining the number of cointegrating vectors are cumbersome. But does this mean that the cointegration principles are of no use? We think not and see possibilities when blending the cointegration approach to the latent variable approaches we used in our paper. This comes close to the suggested approach of using dynamical factorial analysis. Integrating sparsity into the estimation of the cointegration vectors, as proposed in Wilms and Croux (2014), could be another approach to handle high-dimensional data.

The cointegration approach yields predictions and, thus, residuals, which are to be monitored by a control chart. As mentioned by Capizzi, there is a vast literature about the impact of estimation errors on the performance of control charts, with even minor model specifications leading to undesired properties. This is the disadvantage of any model, and the discussion goes back to the previous paragraphs where we state that physics/model based approaches are advantageous for interpretation purposes, but because of limited knowledge of the process under study could be less detailed in describing the data leading to undesired properties of the residuals.

Acknowledgements

We would like to thank Giovanna Capizzi, Marit Schoonhoven and Rob Goedhart again for their valuable comments and suggestions; Guest Editors Ronald Does and Marit Schoonhoven; and the Program Committee of the Stu Hunter Research conference 2015 in Leuven for inviting the first author as a keynote speaker. The first author also acknowledges the Industrial Research Fund of the KU Leuven for their financial support in stimulating the collaboration between academia and industry.

REFERENCES:

Alwan, L. C., Roberts, H. V. (1988). Time Series Modeling for Statistical Process Control. *Journal of Business and Economic Statistics,* 6: 87–95.

Gurden, S. P., Westerhuis, J., Bijlsma, S., Smilde, A. K. (2001). Modelling of spectroscopic batch process data using grey models to incorporate external information. *Journal of Chemometrics,* 15:101-121.

Hubert, M., Reynkens, T., Schmitt, E., Verdonck, T. (2015). Sparse PCA for high-dimensional data with outliers. *Technometrics.* Accepted.

Meng, X., Morris, A. J., Martin, E. B. (2003). On-line monitoring of batch processes using a PARAFAC representation. *Journal of Chemometrics* 17:65-81.

Noorossana, R., Saghaei, A., Amiri, A. (eds.) (2011). *Statistical Analysis of Profile Monitoring.* Hoboken, NJ: John Wiley and Sons.

Rato, T., Schmitt, E., De Ketelaere, B., Hubert, M., Reis, M. (2015). A systematic comparison of statistical process monitoring methods for high-dimensional, time-dependent processes. *AIChE Journal*. Accepted.

Vanhatalo, E., Kulahci, M. (2015). The Effect of Autocorrelation on the Hotelling T^2 Control Chart. *Quality and Reliability Engineering International.* Doi: 10.1002/qre.1717.

Westerhuis, J. A., Derks, E. P. P. A., Hoefsloot, H. C. J., Smilde, A. K. (2007). Grey component analysis. *Journal of Chemometrics* 17:10-11.

Wheeler, D. J. (1991). Shewhart's Chart: Myths, Facts, and Competitors. In *45th Annual Quality Congress Transactions*. American Society for Quality Control: 533-538.

Wilms, I, Croux, C. (2014). Sparse cointegration. *FEB Research Report KBI_1423, KU Leuven.*

Xie, L., Lin, X., Zeng, J. (2013). Shrinking Principal Component Analysis for Enhanced Process Monitoring and Fault Isolation. *Industrial & Engineering Chemistry Research* 52:17475-17486.